PATENT ABSTRACTS OF JAPAN

(11)Publication number : 2002-182685
(43)Date of publication of application : 26.06.2002

---

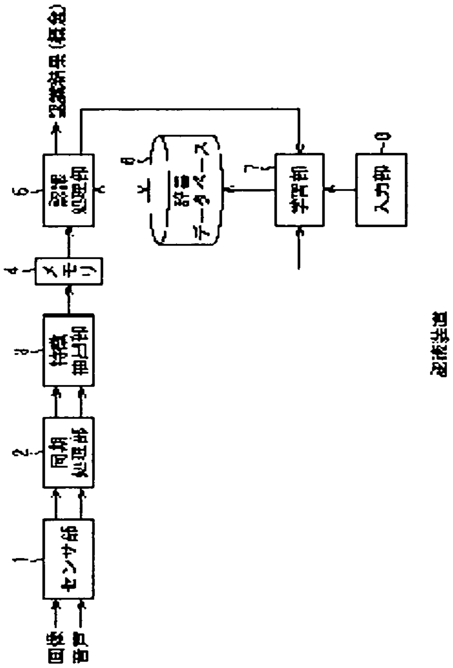
(51)Int.Cl. G10L 15/18
G06T 7/00
G10L 15/00
G10L 15/24

---

(21)Application number : 2000-376911 (71)Applicant : SONY CORP
(22)Date of filing : 12.12.2000 (72)Inventor : NAKATSUKA KOUCHIYO

---

(54) RECOGNIZER AND RECOGNITION SYSTEM, LEARNING SYSTEM AND LEARNING METHOD AS WELL AS RECORDING MEDIUM



(57)Abstract:

PROBLEM TO BE SOLVED: To improve recognition performance.

SOLUTION: A synchronous processing section 2 synchronizes an inputted image and speech and a characteristic extraction section 3 extracts characteristic quantities respectively from the synthesized image and speech and obtains a synthesized characteristic quantity formed by synthesizing the image and speech. A learning section 7 makes learning in accordance with the synthesized characteristic quantity, forms a model dealing with the image and speech indicating the same concept and forms a dictionary which makes the model and the concept information indicating the concept of the image and the speech correspondent to each other. On the other hand, a recognition processing section 5 makes matching by using the synthesized characteristic quantity and the model in the dictionary, thereby recognizing the concept indicated by the input image and speech.

LEGAL STATUS

[Date of request for examination]

[Date of sending the examiner's decision of rejection]

[Kind of final disposal of application other than the examiner's decision of rejection or application converted registration]

[Date of final disposal for application]

[Patent number]

[Date of registration]
[Number of appeal against examiner's decision of rejection]
[Date of requesting appeal against examiner's decision of rejection]
[Date of extinction of right]

* NOTICES *

JPO and NCIPI are not responsible for any
damages caused by the use of this translation.

1.This document has been translated by computer. So the translation may not reflect the original precisely.
2.**** shows the word which can not be translated.
3.In the drawings, any words are not translated.

CLAIMS

[Claim(s)]

[Claim 1] A storage means to memorize the dictionary which matched the image showing an identical concept and an audio model, and the conceptual information showing the concept, An extract means to extract characteristic quantity and to output the synthetic characteristic quantity which compounded the characteristic quantity of the image and voice from said image which synchronized with a synchronous means to synchronize the inputted image and voice, and each voice, Recognition equipment characterized by having a recognition means to recognize the concept which the image inputted by matching using the synthetic characteristic quantity outputted in said extract means and the model in said dictionary and voice express.

[Claim 2] Said synchronous means is recognition equipment according to claim 1 characterized by synchronizing the inputted image and voice by detecting the voice section which are the image section which is the section of the inputted image, and the section of the inputted voice, and normalizing each of said image section and voice section.

[Claim 3] Said synchronous means is recognition equipment according to claim 2 characterized by synchronizing the inputted image and voice by making further in agreement each starting point and terminal point of said image section and voice section which were normalized.

[Claim 4] Said synchronous means is recognition equipment according to claim 2 characterized by synchronizing the inputted image and voice by making the frame of an image and the audio frame in said normalized image section and each voice section correspond further.

[Claim 5] In the recognition approach of performing recognition processing with reference to the dictionary which matched the image showing an identical concept and an audio model, and the conceptual information showing the concept The extract step which extracts characteristic quantity and outputs the synthetic characteristic quantity which compounded the characteristic quantity of the image and voice from said image which synchronized with the synchronous step which synchronizes the inputted image and voice, and each voice, The recognition approach characterized by having the recognition step which recognizes the concept which the image inputted by matching using the synthetic characteristic quantity outputted in said extract step and the model in said dictionary and voice express.

[Claim 6] In the record medium with which the program to which the recognition processing performed with reference to the dictionary which matched the image showing an identical concept and an audio model, and the conceptual information showing the concept is made to carry out to a computer is recorded The extract step which extracts characteristic quantity and outputs the synthetic characteristic quantity which compounded the characteristic quantity of the image and voice from said image which synchronized with the synchronous step which synchronizes the inputted image and voice, and each voice, The record medium characterized by recording the program equipped with the recognition step which recognizes the concept which the image inputted by matching using the synthetic characteristic quantity outputted in said extract step and the model in said dictionary and voice express.

[Claim 7] An extract means to extract characteristic quantity and to output the synthetic characteristic quantity which compounded the characteristic quantity of the image and voice from said image which synchronized with a synchronous means to synchronize the inputted image and voice, and each voice, The model corresponding to the image and voice which generate a model and express an identical concept by learning based on the synthetic characteristic quantity outputted in said extract means, Study equipment characterized by having a study means to generate the dictionary which matched the conceptual information showing the concept of the image and voice.

[Claim 8] Said synchronous means is study equipment according to claim 7 characterized by synchronizing the inputted image and voice by detecting the voice section which are the image section which is the section of the inputted image, and the section of the inputted voice, and normalizing each of said image section and voice section.

[Claim 9] Said synchronous means is study equipment according to claim 8 characterized by synchronizing the inputted image and voice by making further in agreement each starting point and terminal point of said image section and voice section which were normalized.

[Claim 10] Said synchronous means is study equipment according to claim 8 characterized by synchronizing the inputted image and voice by making the frame of an image and the audio frame in said normalized image section and each voice section correspond further.

[Claim 11] The extract step which extracts characteristic quantity and outputs the synthetic characteristic quantity which compounded the characteristic quantity of the image and voice from said image which synchronized with the synchronous step which synchronizes the inputted image and voice, and each voice, The model corresponding to the image and voice which generate a model and express an identical concept by learning based on the synthetic characteristic quantity outputted in said extract step, The study approach characterized by having the study step which generates the dictionary which matched the conceptual information showing the concept of the image and voice.

[Claim 12] In the record medium with which the program to which predetermined study processing is made to carry out to a computer is recorded The extract step which extracts characteristic quantity and outputs the synthetic characteristic quantity which compounded the characteristic quantity of the image and voice from said image which synchronized with the synchronous step which synchronizes the inputted image and voice, and each voice, The model corresponding to the image and voice which generate a model and express an identical concept by learning based on the synthetic characteristic quantity outputted in said extract step, The record medium characterized by recording the program equipped with the study step which generates the dictionary which matched the conceptual information showing the concept of the image and voice.

[Translation done.]

* NOTICES *

JPO and NCIPI are not responsible for any
damages caused by the use of this translation.

1.This document has been translated by computer. So the translation may not reflect the original precisely.
2.**** shows the word which can not be translated.
3.In the drawings, any words are not translated.

DETAILED DESCRIPTION

[Detailed Description of the Invention]
[0001]
[Field of the Invention] Especially this invention relates to a record medium at the recognition equipment and the recognition approach of enabling it to obtain the high recognition engine performance, study equipment and the study approach, and a list, when recognizing the concept from an image and voice about a record medium in recognition equipment and the recognition approach, study equipment and the study approach, and a list.
[0002]
[Description of the Prior Art] In recent years, the robot for entertainment which improvement in the speed of CPU, large capacity-ization of memory, etc. progressed, for example, carried the voice recognition unit is realized by the low price.
[0003] Such a robot does [voice / of a user] speech recognition, and takes various kinds of action based on the recognition result. That is, for example, when a user speaks with a ″hand″, a robot takes action to which an actual dog carries out a hand.
[0004]
[Problem(s) to be Solved by the Invention] If not only the voice as a stimulus given from the outside but an image is processed in a robot, the concept which the voice and image express is recognized and it is made to act in a place based on the recognized concept, it will be expected that the robot which raised entertainment nature more is realizable.
[0005] However, when the voice and the image showing a certain concept are given to a robot for example, since that with which the voice and image synchronized has not become, such an image and voice that do not synchronize are processed, and even if it recognizes the concept which the image and voice express, it is expected that sufficient recognition engine performance is not obtained.
[0006] Moreover, when the voice showing a certain concept and an image are not what synchronized even if it faces study although it is necessary to learn beforehand in order to recognize the concept which an image and voice express as mentioned above, it is expected that sufficient recognition engine performance is not obtained.
[0007] This invention is made in view of such a situation, and when recognizing the concept, it enables it to obtain the high recognition engine performance from an image and voice.
[0008]
[Means for Solving the Problem] A synchronous means to synchronize the image into which the recognition equipment of this invention was inputted, and voice, An extract means to extract characteristic quantity and to output the synthetic characteristic

quantity which compounded the characteristic quantity of the image and voice from the image which synchronized, and each voice, It is characterized by having a recognition means to recognize the concept which the inputted image and voice express by matching using the synthetic characteristic quantity outputted in an extract means, and the model in a dictionary.

[0009] The synchronous step which synchronizes the image into which the recognition approach of this invention was inputted, and voice, The extract step which extracts characteristic quantity and outputs the synthetic characteristic quantity which compounded the characteristic quantity of the image and voice from the image which synchronized, and each voice, It is characterized by having the recognition step which recognizes the concept which the inputted image and voice express by matching using the synthetic characteristic quantity outputted in an extract step, and the model in a dictionary.

[0010] The synchronous step which synchronizes the image into which the 1st record medium of this invention was inputted, and voice, The extract step which extracts characteristic quantity and outputs the synthetic characteristic quantity which compounded the characteristic quantity of the image and voice from the image which synchronized, and each voice, It is characterized by recording the program equipped with the recognition step which recognizes the concept which the inputted image and voice express by matching using the synthetic characteristic quantity outputted in an extract step, and the model in a dictionary.

[0011] A synchronous means to synchronize the image into which the study equipment of this invention was inputted, and voice, An extract means to extract characteristic quantity and to output the synthetic characteristic quantity which compounded the characteristic quantity of the image and voice from the image which synchronized, and each voice, It is characterized by having a study means to generate the dictionary which matched the synthetic characteristic quantity outputted in an extract means, the synthetic characteristic quantity obtained from the image and voice which learn based on synthetic characteristic quantity and express an identical concept, and the conceptual information showing the concept of the image and voice.

[0012] The synchronous step which synchronizes the image into which the study approach of this invention was inputted, and voice, The extract step which extracts characteristic quantity and outputs the synthetic characteristic quantity which compounded the characteristic quantity of the image and voice from the image which synchronized, and each voice, It is characterized by having the study step which generates a model and generates the dictionary which matched the model corresponding to the image and voice showing an identical concept, and the conceptual information showing the concept of the image and voice by learning based on the synthetic characteristic quantity outputted in an extract step.

[0013] The synchronous step which synchronizes the image into which the 2nd record medium of this invention was inputted, and voice, The extract step which extracts characteristic quantity and outputs the synthetic characteristic quantity which compounded the characteristic quantity of the image and voice from the image which synchronized, and each voice, The model corresponding to the image and voice which generate a model and express an identical concept by learning based on the synthetic characteristic quantity outputted in an extract step, It is characterized by recording the program equipped with the study step which generates the dictionary which matched the conceptual information showing the concept of the image and voice.

[0014] The image and voice which were inputted into the recognition equipment of this invention and the recognition approach, and the list in the 1st record medium synchronize, from each of the image which synchronized and voice, characteristic quantity is extracted and the synthetic characteristic quantity which compounded the characteristic quantity of the image and voice is outputted. And the concept which the inputted image and voice express is recognized by matching using the synthetic characteristic quantity and the model in a dictionary.

[0015] The image and voice which were inputted into the study equipment of this invention and the study approach, and the list in the 2nd record medium synchronize, from each of the image which synchronized and voice, characteristic quantity is extracted and the synthetic characteristic quantity which compounded the characteristic quantity of the image and voice is outputted. And by performing study based on the synthetic characteristic quantity, a model is generated and the dictionary which matched the model corresponding to the image and voice showing an identical concept and the conceptual information showing the concept of the image and voice is generated.

[0016]

[Embodiment of the Invention] Drawing 1 shows the example of a configuration of the gestalt of 1 operation of the recognition equipment which applied this invention.

[0017] For example the sensor section 1 consists of a microphone (microphone) which collects and outputs voice (sound), and a video camera which photos and outputs an image, senses the voice and the image as a stimulus which are given from the outside, carries out A/D (Analog Digital) conversion at least, and outputs corresponding voice data and corresponding image data to the synchronous processing section 2.

[0018] About the voice data and the image data which the sensor section 1 outputs, the synchronous processing section 2 gives synchronous processing which is mentioned later, thereby, is a predetermined frame unit and supplies the voice data and the image data which synchronized mutually to the feature-extraction section 3.

[0019] The feature-extraction section 3 processes the voice data supplied from the synchronous processing section 2 in the frame (suitably henceforth voice frame) unit, and extracts an audio feature parameter. Furthermore, the feature-extraction section 3 processes the image data supplied from the synchronous processing section 2 in the frame (suitably henceforth image frame) unit, and extracts the feature parameter of an image. And the feature-extraction section 3 compounds the feature parameter of the voice obtained from the voice frame, and the feature parameter of the image obtained from the image frame corresponding to the voice frame, and supplies the synthetic feature parameter obtained as a result to memory 4.

[0020] In addition, as the synthetic approach of the feature parameter of voice and an image, when the feature parameter of the voice and image consists of vectors, there is the approach of connecting the element (component) of the feature vector of an image with the voice, and constituting one vector etc., for example.

[0021] moreover -- as an audio feature parameter -- MFCC (Mel Frequency Cepstrum Coefficients) and inter-frame [its] -- difference, power, etc. can be used. Furthermore, as a feature parameter of an image, a motion vector, the information showing the configuration of the body displayed on color information, the DCT (Discrete Cosine Transform) multiplier, and the image, etc. can be used.

[0022] Memory 4 stores temporarily the synthetic feature parameter supplied from the feature-extraction section 3.

[0023] Using each model registered into the dictionary of the dictionary database 6, and the synthetic feature parameter memorized by memory 4, the recognition processing section 5 performs matching processing which asks for the score (likelihood) with which a synthetic feature parameter is observed from each model, and asks for the model which matches the synthetic feature parameter most, i.e., the model of for example, the highest score. Based on the highest score, furthermore, the model corresponding to the voice and the image which have been now set as the object of recognition the recognition processing section 5 It judges whether it is finishing [whether it registers with the dictionary of the dictionary database 6, and study] already, and when finishing [study], the conceptual information corresponding to the model of the highest score is searched for, and it outputs as a recognition result of the concept as which the voice into which it was inputted by the sensor section 1, and an image express the conceptual information. Moreover, the recognition processing section 5 requires study of the model from the study section 7, when the model corresponding to the voice and the image which have been now set as the object of recognition is not study ending (when the model corresponding to the voice and the image which have been now set as the object of recognition is not registered into the dictionary of the dictionary database 6).

[0024] In addition, as for a score, likelihood shall become high here, so that the value is large. Moreover, it is possible to use as a score the probability for a synthetic feature parameter to be observed from a model, the similarity (for example, distance of the model and the synthetic feature parameter in a feature space) (Confidence Measure) over the model of a synthetic feature parameter, etc., for example. However, it will be meant that likelihood is so high that the value of a score is small when using the distance of the model and the synthetic feature parameter in a feature space as a score.

[0025] The dictionary database 6 has memorized the dictionary in which the model of the voice obtained by learning using a synthetic feature parameter and an image and the conceptual information showing the concept of the voice and an image are matched and registered. In addition, as a model of voice and an image, a probability model and others, such as HMM (Hidden Markov Model), are employable, for example.

[0026] If the study section 7 has the demand of study from the recognition processing section 5, it will learn the model of voice and an image using the synthetic feature parameter memorized by memory 4. Furthermore, the study section 7 matches the model obtained as a result of study with the voice which the model expresses, and the conceptual information on an image, and registers it into the dictionary of the dictionary database 6.

[0027] The input section 8 consists of keyboards etc., and when inputting the voice corresponding to the model which the study section 7 makes the object of study, and the conceptual information on an image, it is operated by the user. The conceptual information inputted by operating the input section 8 is supplied to the study section 7. In addition, in addition to this, the input section 8 is possible also for constituting from a voice recognition unit etc., and can input conceptual information with voice in this case.

[0028] Next, actuation of the recognition equipment of drawing 1 is explained with reference to the flow chart of drawing 2 .

[0029] The sensor section 1 senses voice and an image and outputs corresponding voice data and corresponding image data to the synchronous processing section 2. The synchronous processing section 2 supplies the voice data for every predetermined voice frame which processed synchronously and synchronized mutually by this about the voice data and the image data which the sensor section 1 outputs, and the image data for every predetermined image frame to the feature-extraction section 3 in step S1.

[0030] In step S2, similarly the feature-extraction section 3 extracts the feature parameter from the image data of the image frame unit from the synchronous processing section 2 while extracting the feature parameter from the voice data of the voice frame unit from the synchronous processing section 2. Furthermore, the feature-extraction section 3 compounds the feature parameter of the voice obtained from each voice frame, and the feature parameter of the image obtained from the image frame corresponding to the voice frame, and memory 4 is made to supply and memorize it as a synthetic feature parameter in step S2.

[0031] And it progresses to step S3, and the recognition processing section 5 performs matching processing, and asks for a score. That is, using each model registered into the dictionary of the dictionary database 6, and the synthetic feature parameter memorized by memory 4, from each model, the recognition processing section 5 asks for the score with which a synthetic feature parameter is observed, further, asks for what has the largest value (the highest score) among the scores of each model, and progresses to step S4.

[0032] In step S4, the recognition processing section 5 judges whether the model of the voice and the image by which the highest score was inputted into the sensor section 1 by whether it is size (beyond a predetermined threshold) from the predetermined threshold epsilon is study ending. In addition to this, the judgment of being finishing [a model / study] can be carried out by input voice in speech recognition here with the application of the technique of judging whether it is the word (unknown word (OOV (Out Of Vocabulary))) which is not registered into a dictionary etc.

[0033] In step S4, when it judges that the highest score is size from the predetermined threshold epsilon, the recognition processing section 5 progresses to step S5 noting that the model of the voice and the image which were inputted into the sensor section 1 is study ending. At step S5, the recognition processing section 5 reads the conceptual information matched with the model of the highest score from the dictionary of the dictionary database 6, outputs it as a recognition result, and ends processing.

[0034] On the other hand, when it judges that the highest score is not size from the predetermined threshold epsilon in step S4, to the study section 7, the recognition processing section 5 requires study of the model, and progresses to step S6 noting that the model of the voice and the image which have been now set as the object of recognition is not study ending.

[0035] At step S6, the study section 7 learns the model of the voice and the image which were inputted into the sensor section 1 using the synthetic feature parameter memorized by memory 4. In addition, when adopting HMM as a model, it is possible to use the re-presuming method of Baum-Welch etc. for study of a model.

[0036] If a model is obtained by study, the study section 7 outputs the message which requires the input of the voice which the model expresses, and the conceptual information on an image from the display or loudspeaker which is not illustrated, when a user operates the input section 8, will wait to input conceptual information and will progress to step S7.

[0037] At step S7, the study section 7 matches with the model obtained as a result of study, and the conceptual information inputted by operating the input section 8, carries out additional registration at the dictionary of the dictionary database 6, and ends processing.

[0038] In addition, although a recognition result is outputted from the recognition processing section 5 in an above-mentioned case or it was made to learn in the study section 7 to it based on the size relation between the highest score and a threshold epsilon The switch which, in addition to this, changes recognition mode and learning mode in equipment is formed, and it can be determined based on actuation of the switch whether a recognition result is outputted from the recognition processing section 5 or it learns in the study section 7.

[0039] Next, the synchronous processing which the synchronous processing section 2 performs in step S1 of drawing 2 is explained.

[0040] For example, when inputting now the voice of "kicking a ball", and the image in the condition have kicked the ball, from the sensor section 1, as it indicates in drawing 3 as the section (voice section) when the voice of "kicking a ball" exists, and the section (image section) when the image in the condition have kicked the ball (image, with which action of kicking a ball is performed) exists, will be synchronized.

[0041] That is, generally the starting point (time of day) Sb of the voice section of the voice of "kicking a ball", and the starting point Mb of the image section of the image in the condition of having kicked the ball are not in agreement, and, generally its terminal point Me of the image section of the image in the condition of having kicked the audio terminal point (time of day) Se and audio ball of the voice section of "kicking a ball" does not correspond, either. Therefore, die-length TM (=Me-Mb) of the image section is not in agreement with die-length TS (=Se-Sb) of the voice section, either.

[0042] Moreover, when changing recognition mode and learning mode with a switch and it is going to learn in learning mode as mentioned above, as shown in drawing 4 , the repeat input of the voice and the image for [the] study may be carried out. Also in this case, the voice section of the voice by which a repeat input is carried out, and the image section of an image will be synchronized like the case in drawing 3 .

[0043] Namely, now, while expressing the voice and the image which are inputted into the i-th as Si and Mi, respectively If the starting point of the voice section of Voice Si, a terminal point, and die length are expressed as B (Si), E (Si), and T (Si), respectively and the starting point of the image section of Image Mi, a terminal point, and die length are expressed as B (Mi), E (Mi), and T (Mi), respectively Generally starting point [of the i-th voice section] B (Si) and starting point [of the image section] B (Mi) are not in agreement, and, generally terminal point E (Si) and its E (Mi) do not correspond. Therefore, generally die-length [of the voice section] T (Si) and die-length [of the image section] T (Mi) are not in agreement.

[0044] Thus, if the starting point of the voice section and the image section, a terminal point, and die length are not in agreement, it will set to the synthetic feature parameter which compounded the feature parameter of voice and an image. Although the section (only either exists) when either of the feature parameters of voice or an image does not exist is generated and both voice and an image are inputted about a certain same concept in this case The recognition engine performance will deteriorate as compared with the case where either voice or the images are used for recognition of the concept, consequently both voice and an image are used for it.

[0045] Then, the synchronous processing section 2 of drawing 1 synchronizes Voice S and Image M which are supplied from the sensor section 1, as shown in drawing 5 .

[0046] That is, the synchronous processing section 2 performs normalization processing which makes in agreement terminal point Se* of the voice section, and terminal point Me* of the image section at other time of day while making in agreement starting point Sb* of the voice section of Voice S, and starting point Mb* of the image section of Image M at a certain time of day as shown in drawing 5 (A). In addition, the die length of the voice section and the die length of the image section will be in agreement by performing normalization.

[0047] The synchronous processing section 2 makes a reference point either starting point Sb* of the audio voice section or starting point Mb* of the image section of Image M, and more specifically makes starting point Sb* of the voice section, and starting point Mb* of the image section in agreement with the reference point. Furthermore, the synchronous processing section 2 makes predetermined time amount, such as 200,400 or 800 mses, conventional-time length, and each die length of the voice section and the image section changes terminal point Se* of the voice section, and terminal point Me* of the image section so that it may be in agreement with conventional-time length. Therefore, terminal point Se* of the voice section and terminal point Me* of the image section will be in agreement in the point that only conventional-time length passed, from a reference point.

[0048] Furthermore, the synchronous processing section 2 performs matching processing which makes each voice frame of the normalized voice section, and each image frame of the image section correspond to one to one as shown in drawing 5 (B), and, thereby, synchronizes voice and an image.

[0049] In addition, since the die length of the voice section or the image section changes, it is necessary to make the number of the voice frame which constitutes the voice section or the image section, or image frames fluctuate according to normalization processing. Moreover, when the time amount length of a voice frame and an image frame differs, it is necessary to make the number of a voice frame or image frames fluctuate also in matching processing. It is possible to perform the change in the number of this voice frame or image frames by interpolation, infanticide, etc.

[0050] although the voice frame and the image frame were matched here at one to one in the above-mentioned case -- a voice frame and an image frame -- one to many or many pairs -- matching with 1 is also possible. That is, it is possible to, match one image frame and L voice frames for example, when the time amount length of an image frame is in agreement L times of the time amount length of a voice frame.

[0051] Next, drawing 6 shows the example of a configuration of the synchronous processing section 2 of drawing 1 .

[0052] The image and voice which the sensor section 1 (drawing 1) outputs are supplied to the section detecting element 11 and memory 12.

[0053] The section detecting element 11 detects the image section of the image supplied there, and the audio voice section, and supplies them to the section normalization section 13 and the synchronization section 14. That is, the section detecting element 11 asks for the motion vector of each image frame whole [for example,], and detects the section when the image frame with a certain amount of motion is continuing as the image section based on the motion vector. Moreover, the section detecting element 11 asks for the power of for example, each voice frame, and detects the section when the voice frame which has a certain amount of power is continuing as the voice section based on the power.

[0054] Memory 12 stores temporarily the image data and voice data which are supplied there.

[0055] The section normalization section 13 reads the voice data of the voice frame which constitutes the voice section supplied from the section detecting element 11, and the image data of the image frame which constitutes the image section similarly supplied from the section detecting element 11 from memory 12. Furthermore, about the voice data of the voice section, and the image data of the image section, the section normalization section 13 performs normalization processing mentioned above, obtains by this the voice data and the image data of the voice (it normalized) section and the image section which made in agreement the starting point, a terminal point, and die length, and supplies them to the synchronization section 14.

[0056] The synchronization section 14 outputs the voice frame and image frame of the voice section and the image section which it normalized to one to one at matching and the feature-extraction section 3 (drawing 1) by performing above-mentioned matching processing about the voice data and the image data from the section normalization section 13.

[0057] Next, with reference to the flow chart of drawing 7 , the synchronous processing section 2 of drawing 6 explains the synchronous processing performed in step S1 of drawing 2 .

[0058] The voice data and the image data which the sensor section 1 (drawing 1) outputs are supplied to the section detecting element 11 and memory 12, and memory 12 stores temporarily the voice data and image data.

[0059] In step S11, the section detecting element 11 detects the image section of the image data from the sensor section 1, and the voice section of voice data, supplies them to the section normalization section 13 and the synchronization section 14, and progresses to step S12.

[0060] At step S12, the section normalization section 13 is reading the voice data of the voice frame which constitutes the voice section from the section detecting element 11, and each image section, and the image data of an image frame from memory 12, and performing normalization processing. The voice data and the image data of the starting point, a terminal point, and the voice section which made die length in agreement and the image section, i.e., the voice section which it normalized and the image section, are obtained, and the synchronization section 14 is supplied.

[0061] In step S13, the synchronization section 14 is performing matching processing which matches the voice frame and image frame of the normalized voice section and the image section of the section normalization section 13 with one to one, it carries out frame synchronization of a voice frame and the image frame, outputs them to the feature-extraction section 3 (drawing 1), and ends synchronous processing.

[0062] As mentioned above, since it was made to synchronize the voice data and the image data showing an identical concept which are inputted from the sensor section 1, as a result of the section when either of the feature parameters of voice or an image does not exist stopping generating in the synthetic feature parameter which compounded the feature parameter obtained from the voice data and image data and performing recognition using such a synthetic feature parameter, the recognition engine performance can be raised.

[0063] Next, hardware can also perform a series of processings mentioned above, and software can also perform. When software performs a series of processings, the program which constitutes the software is installed in a general-purpose computer etc.

[0064] Then, drawing 8 shows the example of a configuration of the gestalt of 1 operation of the computer by which the program which performs a series of processings mentioned above is installed.

[0065] A program is recordable on the hard disk 105 and ROM103 as a record medium which are built in the computer beforehand.

[0066] Or a program is permanently [temporarily or] storable in the removable record media 111, such as a floppy (trademark) disk, CD-ROM (Compact Disc Read Only Memory), MO (Magneto optical) disk, DVD (Digital Versatile Disc), a magnetic disk, and semiconductor memory, again (record). Such a removable record medium 111 can be offered as the so-called software package.

[0067] In addition, it installs in a computer from the removable record medium 111 which was mentioned above, and also from a download site, through the satellite for digital satellite broadcasting services, it transmits to a computer on radio, or a program is transmitted to a computer with a cable through networks, such as LAN (Local Area Network) and the Internet, and by computer, it can receive in the communications department 108 and it can install the program transmitted by making it such on the hard disk 105 to build in.

[0068] The computer contains CPU (Central Processing Unit)102. The input/output interface 110 is connected to CPU102 through the bus 101, and the input section 107 from which CPU102 is constituted from a keyboard, a mouse, a microphone, etc. by the user through an input/output interface 110 will perform the program stored in ROM (Read Only Memory)103 according to it, if a command is inputted by [, such as actuation,] being carried out. Or it is transmitted from the program and satellite with which CPU102 is stored in the hard disk 105 again, or a network, and the program which was received in the communications department 108 and installed on the hard disk 105, or the program which was read from the removable record medium 111 with which the drive 109 was equipped, and was installed on the hard disk 105 is loaded to RAM (Random Access Memory)104, and is performed. Thereby, CPU102 performs processing performed by the configuration of the block diagram according to the flow chart mentioned above processed or mentioned above. and the output from the output section 106 by which CPU102 is constituted from LCD (Liquid CryStal Display), a loudspeaker, etc. through an input/output interface 110 in the processing result if needed or the transmission from the communications department 108 -- record etc. is further carried out to a hard disk 105.

[0069] It is not necessary to necessarily process the processing step which describes the program for making various kinds of processings perform to a computer in this specification here to time series in accordance with the sequence indicated as a flow chart, and it is a juxtaposition thing also including the processing (for example, parallel processing or processing by the object) performed according to an individual.

[0070] Moreover, a program may be processed by the computer of 1 and distributed processing may be carried out by two or more computers. Furthermore, a program may be transmitted to a distant computer and may be executed.

[0071] In addition, the recognition equipment of drawing 1 is applicable to a robot, a voice dialog system, etc. for entertainment. When the recognition equipment of drawing 1 is applied to a robot, a user's utterance (for example, "the ball is rolling" etc.) and the concept which expresses more correctly the event produced around from a corresponding image (for example, image with which signs that the ball was rolling were photoed) are acquired, and it becomes possible to make a robot take exact action etc. Moreover, when the recognition equipment of drawing 1 is applied to a dialog system, it becomes possible to acquire the concept which expresses an intention of a user more correctly and to return an exact response from a user's utterance and gesture, etc.

[0072] Although the concept which they express was recognized from both voice and an image in the gestalt of this operation here, it is also possible to recognize the concept which it expresses from either voice or an image.

[0073] Moreover, it is possible to form the sensor which senses the stimulus from the outside other than an image and voice, such as a microphone and not only a video camera but a pressure sensor, for the information acquired by the sensor to also be included in a synthetic feature parameter, and for it to be made to perform recognition and study in the sensor section 1.

[0074]

[Effect of the Invention] The image and voice which were inputted into the recognition equipment of this invention and the recognition approach, and the list according to the 1st record medium are synchronized, and the synthetic characteristic quantity which extracted characteristic quantity and compounded the characteristic quantity of the image and voice from each of the image which synchronized and voice is obtained. And the concept which the inputted image and voice express is recognized by matching using the synthetic characteristic quantity and the model in a dictionary. Therefore, it becomes possible to raise the recognition engine performance.

[0075] The image and voice which were inputted into the study equipment of this invention and the study approach, and the list according to the 2nd record medium are synchronized, and the synthetic characteristic quantity which extracted characteristic quantity and compounded the characteristic quantity of the image and voice from each of the image which synchronized and voice is obtained. And by learning based on the synthetic characteristic quantity, a model is generated and the dictionary which matched the model corresponding to the image and voice showing an identical concept and the conceptual information showing the concept of the image and voice is generated. Therefore, when recognizing the concept using the dictionary, it becomes possible to raise the recognition engine performance.

[Translation done.]

* NOTICES *

JPO and NCIPI are not responsible for any
damages caused by the use of this translation.

1.This document has been translated by computer. So the translation may not reflect the original precisely.
2.**** shows the word which can not be translated.
3.In the drawings, any words are not translated.

DESCRIPTION OF DRAWINGS

[Brief Description of the Drawings]
[Drawing 1] It is the block diagram showing the example of a configuration of the gestalt of 1 operation of the recognition equipment which applied this invention.
[Drawing 2] It is a flow chart explaining processing of recognition equipment.
[Drawing 3] It is drawing showing signs that voice and an image do not synchronize.
[Drawing 4] It is drawing showing signs that voice and an image do not synchronize.
[Drawing 5] It is drawing showing signs that voice and an image synchronize.
[Drawing 6] It is the block diagram showing the example of a configuration of the synchronous processing section 2.
[Drawing 7] It is a flow chart explaining synchronous processing by the synchronous processing section 2.
[Drawing 8] It is the block diagram showing the example of a configuration of the gestalt of 1 operation of the computer which applied this invention.
[Description of Notations]
1 Sensor Section 2 Synchronous Processing Section, 3 Feature-extraction section 4 Memory, 5 recognition processing section 6 A dictionary database, 7 Study section 8 Input section 11 section detecting element, 12 Memory 13 The section normalization section, 14 Synchronization section 101 A bus, 102 CPU 103 ROM, 104 RAM 105 Hard disk 106 Output section 107 Input section 108 Communications department 109 Drive 110 Input/output interface 111 Removable record medium

---

[Translation done.]

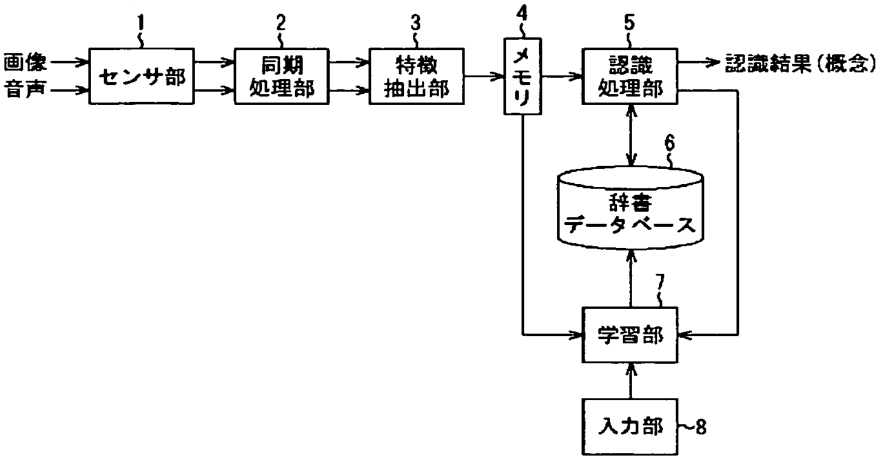
* NOTICES *

JPO and NCIPI are not responsible for any
damages caused by the use of this translation.

1.This document has been translated by computer. So the translation may not reflect the original precisely.
2.**** shows the word which can not be translated.
3.In the drawings, any words are not translated.

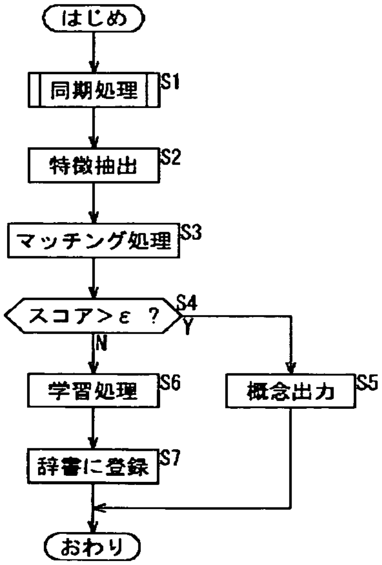
DRAWINGS

[Drawing 1]

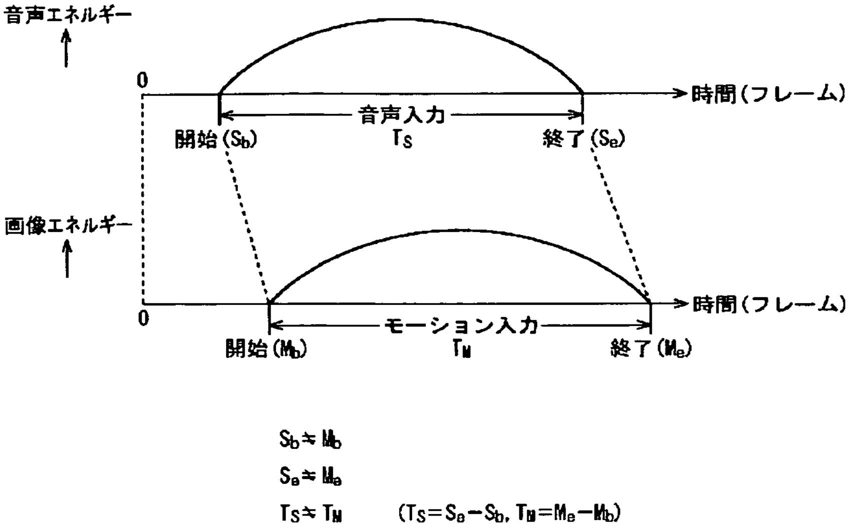

画像
音声
1
センサ部
2
同期
処理部
3
特徴
抽出部
4
メモリ
5
認識
処理部
認識結果（概念）
6
辞書
データベース
7
学習部
入力部
8

認識装置

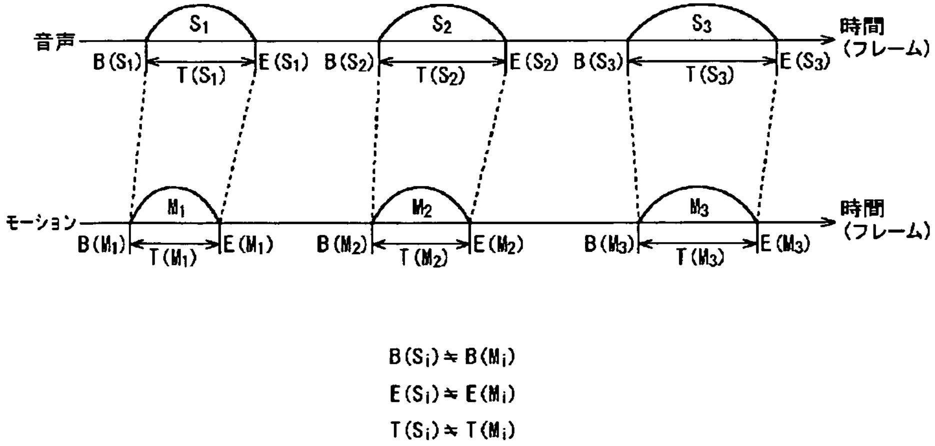
[Drawing 2]


はじめ
同期処理
S1
特徴抽出
S2
マッチング処理
S3
スコア＞ε ?
S4
Y
N
学習処理
S6
概念出力
S5
辞書に登録
S7
おわり

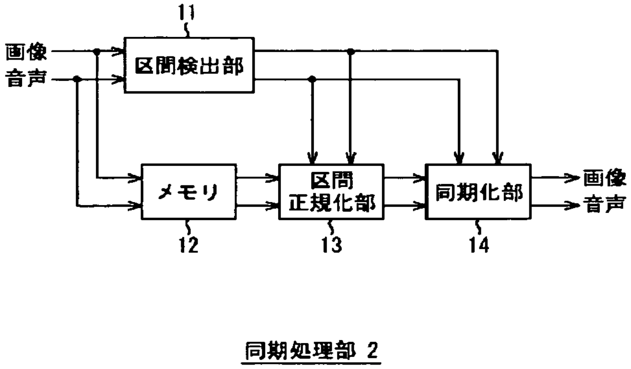
[Drawing 3]



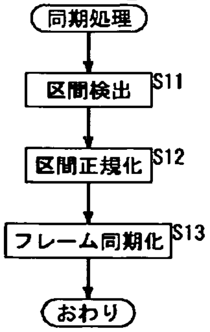
[Drawing 4]



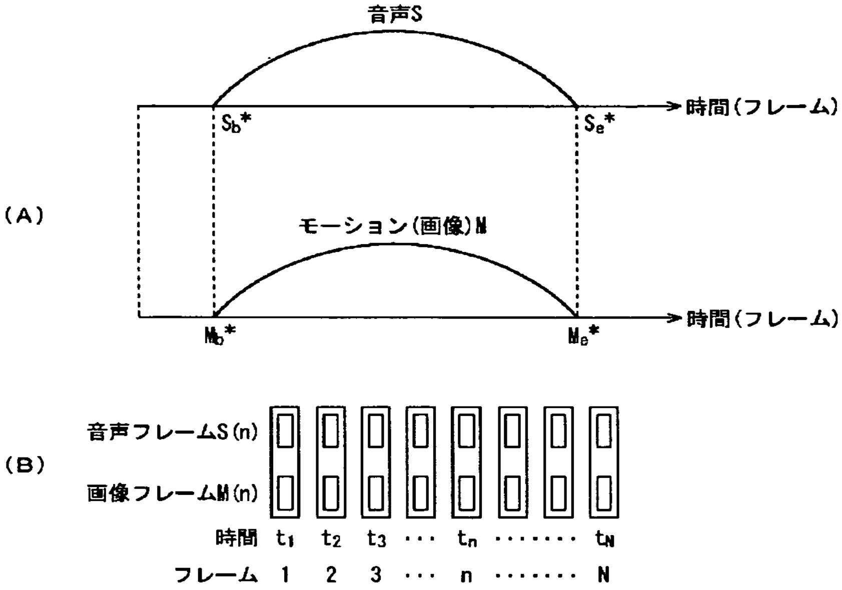
[Drawing 6]


11
画像
音声
区間検出部
メモリ
区間正規化部
同期化部
画像
音声
12
13
14
同期処理部 2

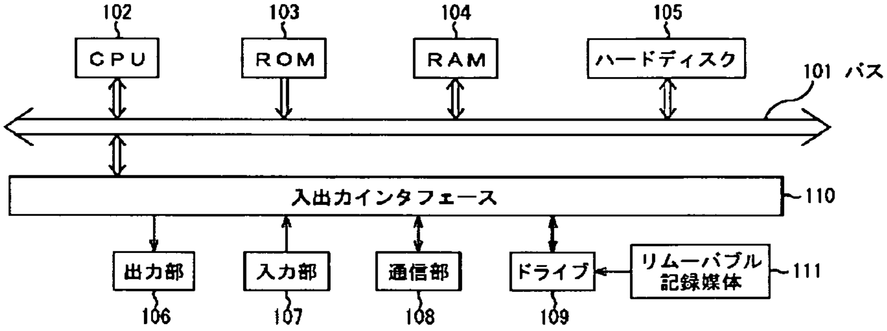
[Drawing 7]


同期処理
区間検出
S11
区間正規化
S12
フレーム同期化
S13
おわり

[Drawing 5]


音声S
時間（フレーム）
S_b*
S_e*
（A）
モーション（画像）M
時間（フレーム）
M_b*
M_e*
音声フレームS(n)
（B）
画像フレームM(n)
時間 t_1 t_2 t_3 ··· t_n ······· t_N
フレーム 1 2 3 ··· n ······· N

[Drawing 8]


102
CPU
103
ROM
104
RAM
105
ハードディスク
101 バス
入出力インタフェース
110
出力部
106
入力部
107
通信部
108
ドライブ
109
リムーバブル記録媒体
111

コンピュータ

[Translation done.]